DOCUMENT RESUME

ED 152 838 TM 007 050

AUTHOR Miller, John K.; Knapp, Thomas R.
TITLE The Importance of Statistical Power in Educational Research. Occasional Paper 13.
INSTITUTION Phi Delta Kappa, Bloomington, Ind.
NOTE 35p.
AVAILABLE FROM Phi Delta Kappa, Eighth and Union, Post Office Box 789, Bloomington, Indiana 47401 ($1.25)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Decision Making; Educational Research; *Hypothesis Testing; Mathematical Models; *Research Design; Risk; Sampling; Statistical Analysis; Statistics; *Tests of Significance
IDENTIFIERS *Statistical Power; Type I Errors; *Type II Errors

ABSTRACT

The testing of research hypotheses is directly comparable to the dichotomous decision-making of medical diagnosis or jury trials--not ill/ill, or innocent/guilty decisions. There are costs in both kinds of error, type I errors of falsely rejecting a null hypothesis or type II errors of falsely rejecting an alternative hypothesis. It is important to consider the power of a statistical test (that is, the likelihood of avoiding type II errors) as well as the significance level (the likelihood of avoiding type I errors) before an experiment is begun. Techniques for estimating the power of a statistical test for a particular experimental design are presented. (CTM)

PHI DELTA KAPPA

EIGHTH AND UNION BOX 789
BLOOMINGTON, INDIANA 47401

PRESENTS AN

OCCASIONAL PAPER

FROM

CENTER ON EVALUATION DEVELOPMENT AND RESEARCH

CEDR

OCCASIONAL PAPER 13

THE IMPORTANCE OF STATISTICAL POWER IN EDUCATIONAL RESEARCH

by

John K. Miller
Thomas R. Knapp
University of Rochester

3 $1.25

PREFACE

The problem of non-significant results in educational experimentation has perplexed researchers for several decades. Two mistakes are made in connection with such findings. The first is the mistake of concluding that in fact there are no differences between the treatments investigated. A non-significant *t* or *F* statistic *does not* mean that there are no differences. Rather, such a statistic states that in this situation we are *unable to reject* the null hypothesis! Failure to reject is not the same as acceptance. The second mistake is one of advanced planning. There have been numerous educational experiments conducted in which *the probability of observing a real difference was practically nil from the start*, a situation that results from the unfortunate choice of sample size, alpha level and desired mean differences. Knapp and Miller have focused on this problem in their treatment of the concept, the power of the test. Educational researchers *can* improve their efforts if they will use the information presented here by those two *as educational experiments are being planned.*

William J. Gephart
Director of Research Services
Phi Delta Kappa

I. INTRODUCTION

A. Human Decision-making

One of the most certain but imperfectly appreciated facts of life is the uncertainty of human knowledge. The conclusions we draw about the world around us are constructed from the data of experience, assembled bit by isolated bit into explanations of how and why things are as they are. We simply do not possess the perceptual equipment for verifying the truth or falsity of any generalization except by the cumulative weight of highly specific observations.

As a consequence the evidence we gather in the effort to generate and verify solutions to problems is rarely, if ever, deterministic. At best the assurance afforded by such evidence is probabilistic. This means, in the absence of certitude, that preference is accorded to solutions most consistent with available evidence and least subject to inexplicable phenomena. Without taking into account every single piece of data relevant to the adequate solution to a problem, one must always anticipate the possibility of finding exceptions, of discovering more comprehensive answers, or even of confirming an altogether different explanation.

By the same token it is often difficult to rule out alternative explanations completely. The amount or quality of evidence available to illuminate the inquiring mind is often inconclusive, providing insufficient basis either for affirming or denying some explanation categorically. And yet we can rarely allow ourselves the luxury of equivocation. We must act and make decisions; and, lacking certainty, if we are to act rationally, we play the odds.

When a physician, for instance, decides upon a diagnosis in a difficult case, he weighs the evidence favoring each possible diagnosis that a patient does or does not suffer from a particular disease, or that the patient is afflicted by one illness rather than another. The danger of error in diagnosis, prescription, and

prognosis is always present, no matter what conclusion the physician ultimately selects. His responsibility is to select the alternative that simultaneously minimizes the prospect of error and maximizes the likelihood of healing or prolonging life. The doctor finds himself, in fact, subject to two quite different but complementary types of error. When there is suspicion of cancer, for example, the physician can arrive at one of two diagnoses – that the patient suffers from the disease or that he does not. The erroneous diagnosis of malignancy could lead to unnecessary endangerment of health and life through medication, radiation therapy, or surgery. On the other hand, the erroneous conclusion that the symptoms do not point to cancer could result in surgically preventable metastasis and premature death for the patient.

Clearly the physician wishes to avoid making either type of error. The trouble is that avoiding one of them is inversely related to avoidance of the other. In other words, the more tests he runs, or the more evidence he demands, or the longer he waits before making a definitive diagnosis of illness and initiates treatment, the greater the danger that the ravages of disease will progress meanwhile to an irreversible stage. The physician will have clung too tenaciously to the *a priori* hypothesis that his patient is not seriously ill. Yet if the doctor relies on purely superficial or ambiguous symptoms, he will readily evade tragic delays in diagnosis, but he may greatly increase the frequency with which he mistakenly treats patients for illnesses they do not have.

The physician's dilemma merely exemplifies, in a context made dramatic by the importance of the decision he must make, the essential peril of the dichotomous decision. Faced with deciding between two mutually exclusive alternatives, supported only by inconclusive evidence, the selection of either alternative could constitute a mistake. The only reasonable solution to this dilemma is to effect an acceptable compromise between the danger of erring in either direction. This can actually be accomplished only by taking account of a number of factors:

1) Explicit identification of both sources of error;
2) Assessment of the severity or cost of each type of error,

3) Comparison of the cost of making one error against the cost of making the other to determine which, if either, is more tolerable;
4) Specification, as precisely as possible, of how much danger making either type of error can be tolerated when a decision is finally made;
5) Control over the inquiry process in a manner that insures adherence to specified error tolerances.

B. Scientific Decision-making

In the natural and behavioral sciences, the testing of research hypotheses is directly comparable to the dichotomous decision-making problem of medical diagnosis. When the scientist systematically pursues the goal of verifying expectations regarding, for instance, the superiority of one instructional method over another, the very nature of his evidence imperils any conclusion he might reach. Should evidence favor the superiority of one method, there always lurks in the background the possibility that uncontrolled, undetected, irrelevant factors are operating to produce only the *appearance* of greater instructional effectiveness. On the other hand, even should evidence fail to support or appear to refute the anticipated outcome, the fault might not lie with the anticipated outcome. It could be the result of defects in the quality of the evidence gathered, or its quantity, or even the manner of obtaining it.

The purpose of scientific inquiry, particularly the investigative techniques we characterize as experimentation, is the acquisition or extension of knowledge by adherence to formalized procedures that minimize the prospect of reaching erroneous conclusions. Kerlinger (1964) points out that the good scientific experiment is designed to let the facts, and only the facts, speak for themselves. Ideally this means that such control is exercised over the sources of error in human judgment that the data will actually support a true hypothesis and will not suggest that an erroneous one is true.

By using highly refined and time-tested methods of inquiry, science seeks to neutralize extraneous or irrelevant factors that

commonly contaminate more casual methods of acquiring knowledge. In pursuit of this ideal, investigators are confronted, more often than not, with the task of gathering, organizing, synthesizing, evaluating, and interpreting an unwieldy array of evidence. Even relatively simple problems would defy experimental solution were it not for the susceptibility of the data of empirical science to quantification. The ability to summarize his observations in the form of measurements places a powerful analytic tool – statistical inference – in the hands of the scientist.

II. STATISTICAL INFERENCE

As we've noted, to produce an acceptable diagnosis the physician must evaluate the symptoms he observes against diagnostic alternatives. The data of science are weighed statistically for their comparative consistency with certain hypotheses, plausible alternative solutions to scientific questions. To accomplish this purpose there is available a wide range of statistical techniques classified in general as hypothesis-testing procedures. These methods enable us to determine the likelihood that the laws of chance alone suffice to account for the occurrence of some event. Statistical analysis would indicate, for example, that the emergence of a seven on ten successive throws of the dice at the gaming table has a very low chance expectancy. Quite logically the gambler would be disinclined to attribute the event to chance and would favor the suspicion that the dice were loaded. Similarly we might discover statistically that the difference in learning rates associated with different instructional methods is too great to readily attribute to fortuitous, non-instructional aspects of pupils' learning experiences. If the findings had been correctly anticipated, then the investigator could claim legitimately to have "confirmed" his hypothesis favoring the effectiveness of one method over the other.

These simple examples point to the essential property of the hypothesis-testing model. Scientific hypotheses are not validated or invalidated in and of themselves. They are confirmed or disconfirmed only in a relative sense, i.e., relative to some explicit

alternative, some competitive hypothesis. Statistically the investigator did *not* compare two teaching methods in the example just cited; he compared two hypotheses. This distinction is a crucial one. Failure to recognize its implications has resulted in a widespread practice of doing only half the job that a scientific experiment is capable of accomplishing.

A. Null and alternative hypotheses

Conventionally, in the course of his investigation, the scientist identifies two mutually exclusive hypotheses: one called the alternative hypothesis (H_1), which he is inclined to believe will stand; and one called the null hypothesis (H_0) which he is inclined to disbelieve. The statistical comparison of the null and alternative hypotheses results in nothing more or less than a simple probability statement – the degree to which a phenomenon, such as the difference observed between two learning rates, would be expected to occur purely by chance. However, the scientist's principal task is to choose between H_0 and H_1.

B. Significance and Type I Error

In relation to that task the statistical probability statement can be interpreted as the likelihood of making an error when the null hypothesis is rejected and the alternative accepted. For example, if this probability is found to be .03 (3%), and the investigator decides to place his trust in the alternative hypothesis that the learning rates do differ as a result of different methods of instruction, he runs a 3% risk of error. Such an error is known as Type I Error.

In order to decide between the null and the alternative, the researcher must have a decision rule, a policy for determining when to side with H_1. He selects some small value, ordinarily 5% or 1%, as the risk of Type I Error that appears, under the circumstances, to be reasonable or tolerable. This risk is commonly called the "level of significance" in statistical jargon and is designated by the Greek letter *alpha* (α). In effect, when the probability of committing Type I Error is found to be as small as the risk an investigator is willing to take, he rejects the null

hypothesis and accepts the alternative hypothesis. That is, he attributes his findings not to the operation of chance but to the systematic influence of factors associated with the alternative hypothesis.

When and if his observations lead the researcher to reject the null hypothesis (H_0) there is, of course, cause for rejoicing. The evidence favors the solution he has formulated to a scientific problem. But let us suppose, as is almost universally true, that the investigator has been satisfied with designing a study that cautiously guards *only* against the commission of Type I Error. Suppose, further, that the outcome of the data analysis does *not* permit the rejection of the null hypothesis, i.e., that the predetermined "level of significance" is not attained, so that adherence to the alternative hypothesis would entail an unacceptable risk of Type I Error. Does this rigorous control against interpreting chance events as experimental effects provide any insurance against attributing real experimental effects to chance? Unfortunately, but emphatically, it does not. Evidence that H_0 is not supported by the data does, indeed, enhance the credibility of H_1. But failure to reject H_0 does not make it credible, or give cause for rejecting the credibility of H_1. It would, for example, be flirting with disaster for a physician to reject the diagnosis of cancer in the face of some, but not all, of the positive clinical signs of the disease. Similarly, failure to find the evidence required to confirm the superiority of one instructional method over another does not refute its superiority. Both these instances exemplify a basic principle of scientific inquiry, indeed of inductive reasoning in general. Failure to confirm any hypothesis does not constitute evidence against it. There is a huge difference between knowing that something *is not true* and *not knowing* that it *is* true.

C. Type II Error and the concept of power

In limiting his susceptibility to Type I Error through specification of a stringent level of statistical significance, the scientist buys protection against false claims to the discovery of new explanations. When the criterion of significance (α) is not satisfied and the alternative hypothesis is unconfirmed, however,

the issue remains essentially unresolved. At best, the investigator may have the option of deferring final judgment until more data become available. At worst, considerations of practical necessity may require him to relinquish his research hypthesis. In either case it remains not only possible but plausible that the alternative (H_1) is true but unsupported by the outcome of the experiment. If H_1 is, in fact, true, an error of the second kind (Type II Error) has been committed. Typically, when results do not satisfy the criterion of significance, the experiment is regarded as a failure. As Campbell and Stanley (1963) point out, experimental failures of this kind are more to be expected than experimental successes. This should be cause for neither surprise nor discouragement. The formulation of acceptable solutions to serious problems is far more difficult than formulating erroneous or incomplete solutions. However, it is unfortunate that *failure to confirm hypotheses* has become equated with *experimental failure*. An experiment truly fails only if it fails to extend the horizons of knowledge.

If one could be reasonably confident that a research hypothesis remains unconfirmed because it is actually false, the goals of science (if not the investigator's goals) would be fulfilled. In other words, an experiment designed to promote simultaneously confirmation or elimination of H_1 never fails. Such a study contributes to knowledge by proposing and supporting an acceptable solution to a problem, or by discovering that the solution advanced is inadequate. Either way the investigation is a productive enterprise.

There is, however, only one way that an experiment can constitute such a *two-pronged* attack on ignorance. It must not only control Type I Error; it must also control Type II Error. First, it is necessary to face squarely the question of the risk involved in making a Type II Error: how serious a mistake is it to ascribe the results of an investigation to the chance hypothesis (H_0), when the investigator's explanation (H_1) is actually an accurate one? What odds does the physician require against diagnosing a malignant tumor (H_1) as some lesser illness (H_0)? Are odds of 9 to 1, allowing a 10% probability of Type II Error, acceptable? Or must he be more cautious, demanding better odds

(e.g., 999 to 1) and a lower risk of error (e.g., 0.1%)? If he is to retain the confidence of his patients, the respect of his colleagues, and his license to practice, his preference had better be for strong safeguards against Type II Error. On the other hand, the consequences of failing to verify the superiority of individualized reading instruction may not be so dire. A 10% risk of abandoning a successful innovation may be acceptable, if the conventional method already in use is known to be reasonably effective.

In the application of statistical hypothesis-testing procedures it is possible to place restraints upon the commission of Type II Error, just as it is possible to limit susceptibility to Type I Error. The risk of making a Type II Error is designated by the Greek letter beta (β), and, like the level of significance (α), it represents a simple probability statement – the probability of failing to reject an erroneous null hypothesis. Statistical *power*, ($1-\beta$) is a correlative concept indicating the probability of rejecting an erroneous null hypothesis in favor of an alternative hypothesis. The physician must ordinarily demand great power of his diagnostic decision in cases where failure to discover a serious illness (Type II Error) might endanger life.

Power must also be high in pharmaceutical experiments in order to evaluate the side effects of medicinal drugs. In Europe the disastrous effects of prematurely marketing the tranquilizer thalidomide is a classic example of the implications of experimental power. Failure to reject the safety of the drug during pregnancy and to discover its damaging effect upon the fetus (H_1) was a costly Type II Error. Had more extensive research been undertaken, requiring uneventful consumption by a larger number of subjects, power would have increased, failure to detect fetal damage might have been avoided, and the drug might never have been erroneously judged safe for distribution.

Though Type II Error in educational and psychological research is rarely accompanied by the danger of such great and irreversible effects, the difference is one of degree rather than kind. The seriousness of the potential error determines *how much* power is necessary. But any experiment designed without specifying a level of power proportional to that necessity is

inherently weak. Let us suppose that the level of power to be required in an educational experiment comparing the effect of televised instruction and programmed instruction has been defined. And let us suppose, for purposes unique to this investigation, that the level is a very demanding one for a learning experiment (Power = .99). What factors influence the attainment of desired power? They are four: 1) the significance level or risk of Type I Error that is deemed tolerable; 2) the difference between the two treatment-population means on the learning criterion; 3) the variance in the population on the learning criterion; and 4) the number of subjects in each of the treatment groups. The selection and attainment of some preferred level of power requires that each of these four values be specified in advance.

The choice of significance level is, or should be, a prudential judgment based on assessment of the consequences of committing a Type I Error. The more severe the impact of erroneously rejecting the null, the more rigorous the level of significance must be.

The smallest difference between means that would be of interest to the investigator must be selected. This represents the precise definition of a very specific alternative hypothesis. It should be noted that the exercise of control over statistical power precludes the formulation of the more conventional, non-specific alternative hypothesis. Neither the so-called "two-tailed" alternative (i.e., that the mean difference does not equal zero), nor a "one-tailed" hypothesis (i.e., that the mean difference favors one treatment over the other) is sufficient. All other things being equal, the power of the decision to accept or reject the null hypothesis is directly affected by variation in the difference between means.

Power is sensitive, however, not only to differences between means, but to variation in performance among individuals as well. The wider the range of performance to be found on the experimental criterion variable, the greater the danger of committing Type II Error. It is therefore necessary, by rational or empirical means, to anticipate the extent of individual differences

(the variance) characteristic of the criterion measure.

Finally, variations in sample size also affect the power of inference. However, when desired levels of power and significance have been pre-selected, and when a mean difference worthy of the investigator's interest has been defined, and when a reasonable variance estimate is available, then the number of experimental subjects needed is no longer free to vary. In fact, the most feasible method for bringing Type II Error under control is to calculate and select exactly that number of subjects required to satisfy those conditions.

III. AN ANALOGY AND AN APPLICATION

In a jury trial the guiding (null) hypothesis is that the defendant is innocent until judged guilty.* If at the conclusion of a trial the defendant is not judged to be guilty, i.e. the null hypothesis is not rejected, there are two possible explanations: (1) the defendant is in fact innocent, i.e. the null hypothesis is true; or (2) the defendant is in fact guilty but an error (Type II) in judgment has been made, i.e. a false null hypothesis has been retained, because of insufficient or inconclusive evidence.

Similarly, in experimental educational research the guiding (null) hypothesis is that there is no difference between (among) the experimental treatments. If at the conclusion of the experiment one of the treatments is not judged to be better than the other(s), i.e. the null hypothesis is not rejected, there are two possible explanations: (1) the treatments are in fact equally effective, i.e. the null hypothesis is true; or (2) one of the treatments is in fact better than the other(s) but an error (Type II) in judgment has been made, i.e. a false null hypothesis has been retained, because of insufficient evidence (sample size too small, reliability of the dependent variable too low, etc.).

*The expression, "innocent until *proven* guilty" is used more often, but "proof" in any absolute sense is impossible to establish in a jury trial or in experimental educational research.

What is the relevance of power to both of these situations? For the jury trial, society, through its representatives (the presecuting and defense attorneys, the judge, the jury) must decide, *before the trial begins*, what risks it is willing to assume as far as both Type I Error (declaring as guilty an innocent defendant) and Type II Error (declaring as innocent a guilty defendant) are concerned, and act accordingly. It must choose a decision-making procedure (trial) with appropriate power (by considering a large amount of evidence in a long trial, for example, if both errors are very serious and equally serious, and if a fine discrimination between guilt and innocence must be made). For experimental educational research, the scholarly community, through *its* representative (the researcher) must decide, *before the experiment begins*, what risks *it* is willing to assume as far as both Type I Error (declaring as different equally effective treatments) and Type II Error (declaring as equal differentially effective treatments), and *also* act accordingly. It must also choose a decision-making procedure (significance test) with appropriate power (by drawing a large sample of subjects and using precise measuring instruments, for example, if both errors are serious, and if a rather fine discrimination between equal effectiveness and superiority must be made).

Table 1 explores, step by step, the parallels between a jury trial and an experimental educational research investigation into the problem of the relative effectiveness of two teaching methods, e.g. televised instruction and programmed instruction, for a particular unit of a course in, say, secondary school (tenth grade) biology.

Table 1: Analogy Between Jury Trial and Experimental Research

	Jury Trial	Experimental Research
Null hypothesis:	The defendant is innocent.	On the average, televised instruction and programmed instruction are equally effective: i.e. $\mu_T = \mu_b$ or $\mu_T - \mu_b = 0$
Alternative hypothesis:	The defendant is guilty.	On the average, televised instruction is ten points more effective than programmed instruction; i.e. $\mu_T - \mu_b = 10$
Risks Type I (α) and Type II (β) error:	Both errors, viz. deciding that an innocent person is guilty (Type I Error) and, deciding that a guilty person is innocent (Type II Error), are very serious and equally serious.* Therefore both α and β should be equally small.	The cost of making a Type I Error, viz. deciding that one method is better than the other when in fact they are equally effective, is not substantial, but the cost of making a Type II Error is, viz. deciding that these methods are equally effective when in fact televised instruction is actually better. Therefore β should be smaller than α, say .01 as compared to .05, necessitating the choice of a significance test (one-tailed) with power = .99 (=1 - β).
Selection of sample size:	The only way to keep α and β equally small is to have a large N, i.e. to collect a large amount of evidence and conduct a long trial.	See Figure 1 and associated calculations in text.

*Scheff (1963) and most present-day liberals argue that deciding that an innocent person is guilty is a much more serious error (Type I). We find it difficult to make such a value judgment.

Statistic	The difference between the evidence in support of the defendant's guilt and the evidence in support of his innocence.	The difference between the mean scores on the TOUS for two independent randomly assigned samples.
Decision rule:	If all jurors agree that guilt has been established, reject H_0; otherwise, retain H_0. (Such a decision is often based, consciously or unconsciously, on the probability of the association of the defendant with the conditions of the crime.*)	If the difference between the sample means is greater than 4.14, reject H_0; otherwise, retain H_0. (The probability of the difference greater than 4.14 is less than .05 if H_0 is true and is greater than .99 if H_1 is true.)

*This point is illustrated in the following example taken from Kingston (1965a):

> Consider . . . the following hypothetical case. An International courier, carrying a brief case cuffed to him, in which was carried a considerable amount of money and some secret documents, was murdered in London. The brief case was ripped open and the contents taken. The only evidence found by the investigating authorities was a latent fingerprint on the brief case that could only have been left by the perpetrator. It is calculated that the probability of the ridge pattern shown by the latent print occurring by chance on any one person is about $1/(3X10^9)$. The latent print is filed with an international police record office after an otherwise unsuccessful investigation. Every fingerprint filed coming to the attention of this office is compared with the latent print. One year after the crime, such a routine comparison turns up a fingerprint card which contains an impression area matching the latent print. This card was made from a person applying for a job in Washington D.C. He is immediately arrested and charged with the crime. Considering that the above probability calculation is accurate, what is the chance that the above accusation is in error, where the latent print is the only evidence that can be used?

The principal problem is one of solving for the N (assuming equal sample sizes for the two groups) which is such that the point marked "X" (Figure 1) cuts off 5% (=α) of the area in the right-hand tail of the null distribution and 1% of the area in the left-hand tail of the alternative distribution, i.e. (assuming normality for both sampling distributions) $z_{null} = +1.645$ and $z_{alt.} = -2.327$. The standard error for each distribution (assuming equal variances) is given by

$$\sqrt{\frac{\sigma^2}{N} + \frac{\sigma^2}{N}} = \sqrt{\frac{2\sigma^2}{N}} = \frac{1.414\sigma}{\sqrt{N}}$$. The required calculations proceed as follows:

null distribution $\frac{X - 0}{\frac{1.414\sigma}{\sqrt{N}}} = 1.645 \rightarrow X = \frac{2.326\sigma}{\sqrt{N}}$ (1)

alternative distribution $\frac{X - 10}{\frac{1.414\sigma}{\sqrt{N}}} = -2.327 \rightarrow X = \frac{-3.290\sigma}{\sqrt{N}} + 10$ (2)

Equating the two expressions for X we have:

$$\frac{-3.290\sigma}{\sqrt{N}} + 10 = \frac{2.326\sigma}{\sqrt{N}}$$

or $-3.290\sigma + 10\sqrt{N} = 2.326\sigma$

or $10\sqrt{N} = 5.616\sigma$

or $\sqrt{N} = .562\sigma$

i.e. $N = (.562\sigma)^2 = 316\sigma^2$

Suppose that the dependent variable to be measured at the conclusion of the experiment is the performance of each subject on the Test on Understanding Science (TOUS). The manual for that test (Cooley and Klopfer, 1961) contains the information that for a normative sample of 1055 tenth-graders the standard deviation is 7.66 (the test contains 60 multiple-choice items). Substituting this value into the expression for N we have:

$$N = .316\ (7.66)^2 = .316\ (58.68)$$
$$= 18.54 \text{ or approximately } 19 \text{ subjects per treatment group*}$$

The value of X (the "critical" *obtained* difference between the means for the two *samples*) is found by substituting in equation (1), or equation (2), as follows:

$$X = \frac{2.326(7.66)}{\sqrt{19}}$$
$$= 4.14$$

$$\text{or} \quad X = \frac{-3.290(7.66)}{\sqrt{37}} + 10$$
$$= 4.14$$

Thus, in order to reject the null hypothesis at the .05 level of significance and the .99 power level when the investigator hypothesizes a ten-point difference between means on a variable typified by a standard deviation of 7.66, the optimal number of

*The word "subjects" is used here in its most general sense, i.e. *observations*. The experimental unit may be an individual person, a classroom, a school, or what-have-you. Furthermore, the subjects may be "run" one-at-a-time or as an interacting group. Neither of these matters affects the statistical determination of N but both are of critical importance in the interpretation of the data and the generalizability of the results.

subjects in each of the groups is 19. Moreover, when these conditions are satisfied, the null hypothesis would, in fact, be rejected for an observed difference of 4.14 or larger. In other words, if the difference between the two methods were truly 10 criterion points, an observed difference of 4.14 would be sufficient to prevent the Type I Error rate from exceeding 5% and to prevent the Type II Error rate from exceeding 1%.

Even if the alternative hypothesis is non-specific (e.g, $\mu_T - \mu_p \neq 0$, $\mu_T - \mu_p > 0$, or $\mu_T - \mu_p < 0$) and/or the population standard deviation is unknown, one can solve for N by using an approximation procedure due to Cohen (1969) which is also found in the recent elementary statistics text prepared by Welkowitz, Ewen, and Cohen (1971). All the researcher need do is specify a relative "effect size" (the ratio of a meaningful difference to the standard deviation of the dependent variable) which would be "too good to miss" and use the tables which Cohen has compiled to find the N that does the job. For our example, if we were interested in detecting a "large" effect (Cohen's $\gamma = .80$) we would find (Welkowitz et al., 1971, p. 199) that:

$$N = 2\left(\frac{\delta}{\gamma}\right)^2 \quad (\delta = 3.97 \text{ for } \alpha = .05, \text{ one-tailed, and power} = .99)^*$$

$$= 2\left(\frac{3.97}{.80}\right)^2$$

$$= 2\,(4.96)^2$$

$$= 2\,(24.60)$$

$$= 49.20 \text{ or approximately 49 subjects per treatment group}$$

For a "small" effect, $\gamma = .20$ and N would be very large (about 787 per group); for a "medium" effect, $\gamma = .50$ and N would be about 126 per group.

*δ is a measure which combines the significance level and the power level.

Finally, in the unfortunate event that one is "stuck" with a "grab sample" of subjects which may be more or less than optimal in number, one can at least determine the power for the "available N" and discover if experimental conditions actually provide reasonable protection against Type II Error. Suppose that the total pool of subjects consists of 30 students (15 per treatment group) and we want to test the null hypothesis against the same specific alternative hypothesis of a ten-point difference in favor of televised instruction, with $\sigma = 7.66$ and $\alpha = .05$. From equation (1) on page 14 we have

$$X = \frac{2.326(7.66)}{\sqrt{15}}$$

$$= 4.60$$

Substituting this in the initial formulation for equation (2) from page 14:

$$\frac{4.60 - 10}{\frac{1.414(7.66)}{\sqrt{15}}} = z_{alt} = -1.931$$

A z of −1.931 cuts off 3% of the left-hand tail of the normal distribution. Therefore $\beta = .03$ and power = .97.

IV. WHERE HAS ALL THE POWER GONE?

The two most frequently referenced statistics books in the contemporary experimental research literature of education and psychology are those by Winer (1962) and Hays (1963). Both texts contain excellent discussions of very simple procedures for determining the appropriate sample size (N) for a desired level of statistical power ($1-\beta$), given a specific alternative hypothesis ($H1_J$) of interest and significance level (α) chosen to test the null hypothesis ($H1_0$). As Chandler (1957) so aptly pointed out, power is "...the basic concept responsible for one's employing statistical tests as a basis for taking action on an H(ypothesis)." If power were of no consequence we could adopt the arbitrary convention

of rejecting a randomly selected 5% of all null hypotheses. However, aside from an occasional rationalization of failure to confirm some pet alternative to the null, consideration of power is conspicuous by its absence from research discussions. Investigators continue to carry out mean-difference tests on as many handy subjects as they can lay their hands on and feasibly accommodate.

Why has this apparently crucial aspect of the hypothesis-testing approach to statistical inference been so consistently ignored?

1. The cynic would claim that rejecting H_0, whether it be true or false, has become a matter of personal survival. Power must succumb to more important considerations such as finding significant differences, breaking into print, and obtaining salary increases and promotions. We reject this accusation, perhaps naively, since we take a more optimistic view of the dedication of behavioral scientists to the advancement of knowledge.

2. The defeatist would argue that the combination of high power and a stringent significance level requires a prohibitively large N. For any H_1 that differs only slightly from the null, this is true. Yet the literature contains examples of research studies where investigators have used *more* subjects than would be necessary to strike an optimal balance among power, statistical significance, and meaningful differences. Such studies ignore the implications that power holds for sample size, permit variations in sample size to govern the statistical decision, and, as a consequence, sacrifice relevance on the altar of significance.

3. The scholar would say that most investigators fail to appreciate the hypothesis-testing model in general and the matter of power in particular. Witness the virtual boycott of power

*An analysis of ten of the most popular books on research methods reveals that Best, Borg, Kerlinger, Mouly, Sax, Travers, and Van Dalen devote no space; Wiersma devotes less than a page; Helmstadter allots three pages; and Fox gives fifteen pages to the consideration of power.

concepts by texts devoted to the principles of research design and the conduct of scientific inquiry,* despite the clarity of presentations by Winer and Hays, and despite the periodic emergence of concern about the susceptibility of decision-making behavior to Type II Error (e.g., Cohen, 1962, Kennedy, 1970). Failure to comprehend the simple notions associated with power and sample-size determination is not at all consistent with widespread evidence of increasing sophistication in the complexities of analysis of covariance, multiple discriminant analysis, etc. Many instructors, it is true, hesitate to talk about power in introductory statistics courses,* fearing that students might find it too dificult. Instructors in more advanced courses** may assume, on the other hand, that the notion of power is already part of their students' statistical repertoires.

4. The rigid empiricist would adopt the view that power is itself an object of inquiry rather than a legitimate tool of efficient experimental design. He would label "unscientific" the *a priori* specification of a difference that satisfies some criterion of practical significance. He might even be perturbed by the use of an estimate of the common population variance drawn second-hand from existing information sources. A pilot study at the very least, perhaps even an extensive preliminary sampling survey, are,

*The number of pages devoted to power in ten of the most popular introductory statistics texts is as follows: Blommers and Lindquist, 14; Downie and Heath, 0; Edwards - *Statistical Analysis*, 2; Fergusoh, 0, Games and Klare, 0; Garrett, 0; Guilford - *Fundamental Statistics in Psychology and Education*, 14; Popham, 1; Tate, 1, and Walker and Lev - *Elementary Statistical Methods*, 8.

**In surveying ten texts which can best be described as either advanced or intermediate in difficulty, power received page allottments as follows: Cooley and Lohnes, 0; Edwards - *Experimental Design in Psychological Research*, 5; Glass and Stanley, 6; Hays, 11; Lindquist - *Design and Analysis of Experiments in Psychology and Education*, 0; Marascuilo, 23; McNemar, 2; Walker and Lev - *Statistical Inference*, 7, Wert, Neidt, and Ahmann, 0; and Winer, 4.

under most circumstances, an investigator's best recourse in the search for accurate parameter estimates. Yet experience suggests that the choice of a logical, realistic alternative hypothesis and a reasonable estimate of the population variance can often be based on other considerations. This matter is pursued further in our concluding remarks.

5. The Bayesian would suggest that the people who use the hypothesis-testing model have no real faith in its applicability to behavioral research. Consequently, they do a sloppy job of implementing it. This appears, in some respects, the most convincing and insightful conjecture. The hesitance with which implications and conclusions are extracted from observed results betrays a woeful lack of trust either in the research results themselves or in the strength of the hypothesis-testing model as a framework for inquiry. The reluctance of the practicing educator or psychologist to take seriously the implications of experimental findings further mirrors the researcher's own skepticism. The simple fact of the matter is that the hypothesis-testing model is a dichotomous decision-making model. Its user goes "all the way" with it, or he cripples it. Essential components of the model include: a) *two* explicitly defined hypotheses (H_0 *and* H_1); b) explicit knowledge or reasonable expectations regarding sampling distributions; and c) commitment to accept and act upon the conclusion mandated by the statistical decision. Perhaps there are very few important problems in education and psychology which lend themselves to rigorous focus on the dichotomy: "reject H_0 or reject H_1." We doubt it. But if such is the case, then pretense should be abandoned in favor of other models.

V. CONCLUSION

What is the prescription for salvaging the hypothesis-testing model in those situations for which it is the appropriate procedure? At the very least, it would appear, the power calculation must become as integral a concern in the experimental process as the significance test itself. The complementarity of Type I and II errors should be sufficient cause for at least acknowledging the problem of power far more frequently. It is,

after all, the scientist's responsibility to minimize error — any error. The prevalent preoccupation with the avoidance of Type I Error bespeaks a commendable concern for the integrity of knowledge by subjecting H_1 to rigorous tests against H_0. Yet the cause of science may be prejudiced far more gravely in the long run by the erroneous, and perhaps permanent, abandonment of a true H_1. By its very nature the incorrect rejection of H_0 invites ultimate exposure. Type II Error, however, is more likely to escape detection. Nonetheless, we would not advocate improving power at the expense of stringent significance levels. The ideal experiment is both powerful and rigorous.

It is possible to achieve any desired power level in conjunction with a specified significance level by controlling the number of experimental observations to be taken. For mean-difference research, the control of power through sample size determination depends upon the specification of a difference in magnitude that is *meaningful*, together with an estimate of the common population variance for the dependent variable.

There appears to be a curious reluctance in the behavioral sciences to take responsibility for specifying how large a difference will be regarded as a meaningful difference. Perhaps the attitude persists because concern for meaningful differences hints at a utilitarian view of science, or because a *a priori* specification of valued outcomes appears wanting in scientific objectivity. Whatever the reason, however, the test of statistical significance has emerged as the major arbiter of the value attached to phenomena observed in behavioral research. Yet within both basic and applied fields the interest value of experimental outcomes surely varies (to some degree at least) with the absolute magnitude of the differences observed. Rare trivia are no less trivial than the garden variety.

The abundance of anecdotes about the enhancement of inconsequential outcomes by the artifice of increasing sample size attests to the need for non-statistical criteria for evaluating the importance of research results. The susceptibility of statistical inference to the vagaries of sample size is, in itself, grounds for dismissing statistical significance as the sole criterion of substan-

tive importance. Presumably the competent researcher does have at least an appreciation of his field or discipline sufficient to differentiate trivial from non-trivial differences. If he can not evaluate a hypothetical difference in the light of his theoretical constructs or in relation to prevailing professional practice, there is faint hope for his ability to interpret an observed difference. The real utility of the significance test is the assurance it can provide that evidence supporting some interesting alternative to H_0 is, at the same time, a non-chance event.

Determining in advance the order of magnitude that would distinguish an uninspiring difference from an exciting one offers two important advantages. Besides facilitating advance control over Type II Error tolerance, the procedure effectively excludes from the region of rejection differences too small to be of interest. The net effect is reasonable assurance of rejecting the null when the observed difference is as large as the appropriate predetermined value and avoidance of the embarrassing responsibility for continuing plausible conclusions and implications from inconsequential, but statistically significant, results.

Estimation of the common population variance is an empirical problem – one that is more tractable than is generally believed. Many studies employ instruments about which a great deal is already known. Large numbers of published tests that have achieved popularity as research instruments provide normative data based on large and reasonably representative standardization samples. Many other research studies use unpublished measures that have been used before and for which there is available a rough approximation to the population variance. And, finally, responsible research which employs newly designed instruments will make provision for the acquisition of reliability and validity data that include the variance estimate required for the power calculations. It would be unfortunate to ignore such rich sources of information simply because they were not produced with the explicit intent of facilitating the control or evaluation of statistical power or, even worse, because they were not a product of the investigator's own efforts. Acknowledging dependency on external sources of relevant statistical input might eventually contribute to diminishing the fragmentation so characteristic of

research products in the behavioral sciences. Present custom, characterized by emphasis on the *uniqueness* of an author's contribution, tends to weaken, rather than strengthen, the continuity of science and the relevance of new discoveries.

Given the importance of statistical power to the utility and integrity of difference testing as a decisive contributor to the advancement of knowledge, there is no room for equivocation. Estimating the common population variance, specifying a difference that can be regarded as meaningfully large, selecting a power level that enforces reasonable restraints upon Type II Error, and calculating the sample size required to satisfy the power specification should be as routine as the selection of a significance level. Moreover, like the selection of significance level, these tasks should precede the collection of a single piece of data.

But what if the investigator can honestly claim that for his study there is no intuitive, rational, or empirical basis for stipulating some difference "too good to miss"? The Utopian age of ratio scaling forecast for psychological measures by Wright (1968) has yet to dawn. As a result, the cautious investigator may be deterred by the arbitrariness of his scale from attributing substantive meaning to specific differences or to variance estimates. The solution to this problem rests with the concept of "effect size" exploited in Cohen's (1969) magnificent contribution to the statistical literature, a reference work devoted exclusively to power, complete with tables for sample size determination for virtually every commonly used test of significance. All one need do is specify the *relative* order of magnitude of a meaningful difference, choose the preferred significance and power levels, and read off the tabled sample size. Not only is the notion of effect size an intuitively pleasing one (e.g. an anticipated difference equal to half the pooled samples' standard deviation for the t-test employed with independent samples); it is a practical one. Cohen even provides the user with a rational basis for identifying for each statistic effect sizes that may be described as small, medium, and large in the light of results typically associated in the behavioral sciences with weak, moderate, and potent experimental treatments.

What practical implications does advance determination of sample size have for the design and conduct of research? If the investigator discovers that the sample required is too large to be accommodated by available resources, he has the opportunity to avail himself of a number of *reasonable* alternatives. If both the power level and the significance level chosen are really crucial, the study may be abandoned or deferred until a sufficient number of subjects can be mustered. It may be possible, on the other hand, to effect an acceptable compromise between power and significance by tolerating an increased probability of Type I Error in order to bring Type II Error under sensible control. Or the investigator may take a calculated risk and proceed with the proposed study, fully aware of the extent to which his procedure is less than optimal.

Suppose, however, that desired levels of power and significance do not drain the supply of available subjects. There is, obviously, no scientific virtue in the extravagance of fruitlessly large samples, particularly if smaller groups might permit notable improvements upon the original research design by permitting inclusion of additional experimental treatments or the investigation of interesting interactions.

Can experimental behavioral science really afford the dubious luxury of continuing to blunder upon significance, subject to the fortuitous, coincidental attainment of sufficient statistical power?

ANNOTATED BIBLIOGRAPHY

Campbell, D. T., and Stanley, J. C. "Experimental and Quasi-Experimental Designs for Research." In Gage, N. L. (ed). *Handbook of Research on Teaching.* Chicago: Rand McNally and Company, 1963.

This already-classic chapter on the non-mathematical aspects of experimental design is also available as a separate paperback with a slightly altered title (same publisher, 1966).

Chandler, R. "The Statistical Concepts of Confidence and Significance." *Psychological Bulletin,* 1957, 54, 429-30.

This very brief two-page article clarifies the distinction between significance (a term associated with the likelihood of getting a difference between a particular statistic and a parameter, in *hypothesis testing*) and confidence (a term associated with the likelihood that a particular interval around a statistic covers the parameter, in *interval estimation*). The two terms are often confused with one another in the literature pertaining to inferential statistics.

Cohen, J. "The Statistical Power of Abnormal - Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology,* 1962, 65, 145-153.

This article was the first of many efforts by Cohen to point out the neglect of statistical power in behavioral research. The first part of the article is a summary of the basic concepts involved in power analysis, with examples. The second part is devoted to a critique of 78 articles which appeared in volume 61 (1960) of the *Journal of Abnormal and Social Psychology,* focusing on the question: "What kind of chance did these investigators have of rejecting false null hypotheses?"

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press, 1969.

A textbook devoted entirely to the concept of power, complete with formulas and tables for determining either the sample size required for a given power or the power associated with a given sample size. All of the commonly-encountered significance tests – single sample mean, difference between two sample means (independent or correlated), sample correlation coefficient, etc. – are treated.

Cooley, W. W., and Klopfer, L. E. *Manual for Administering, Scoring, and Interpreting Scores – Test on Understanding Science, Form W.* Princeton, New Jersey: Educational Testing Service, 1961.

The data provided in this manual (means, standard deviations, etc.) are illustrative of the kinds of information a researcher might need if he were carrying out an experimental educational investigation for which performance on this test were the principal dependent variable.

Hays, W. L. *Statistics for Psychologists.* New York: Holt, Rinehart, and Winston, 1963.

The very popular textbook used in many educational and psychological statistics courses throughout the country. It has an excellent section on power (pp. 269-280).

Kennedy, J. J. "A Significant Difference Can Still be Significant." *Educational Researcher,* October 1970, 2, 7-9.

One of many reactions to an article by Coats, in a previous issue of the same journal, objecting to the study of inferential statistics. Kennedy suggests, among other things, the tailoring of sample size to the conditions of the experiment by employing statistical power analysis.

Kerlinger, F. N. *Foundations of Behavioral Research.* New York: Holt, Rinehart, and Winston, 1964.

This popular text in research methods does not treat power as such but does bring to the researcher's attention some of the basic issues involved in statistical inference.

Kingston, Charles R. "Applications of Probability Theory in Criminalistics." *Journal of the American Statistical Association*, 1965, 60, 70-80. (a)

Kingston, Charles R. "Applications of Probability Theory in Criminalistics – II." *Journal of the American Statistical Association*, 1965, 60, 1028 - 1034. (b)

A pair of articles concerned with models for evaluating physical evidence for criminal trials. Power is not explicitly considered, but the probabilistic basis on which it rests is treated in detail.

Scheff, T. J. "Decision Rules, Types of Error, and Their Consequences in Medical Diagnosis." *Behavioral Science*, 1963, 8, 97-107.

As the title indicates, this article explores the relative consequences of Type I Error ("judging a well person sick") and Type II Error ("judging a sick person well") in the field of medicine. The author questions the usually – unwritten assumption that Type II Errors are more serious in this context, whereas he supports the notion in the field of law that it is worse to convict an innocent person than to let a guilty one go free.

Welkowitz, J., Ewen, R. B., and Cohen, J. *Introductory Statistics for the Behavioral Sciences*. New York: Academic Press, 1971.

This very new textbook is one of the few introductory texts which contains more than a page or two on power. In Chapter 13 the authors treat both sample size and power determination for the four most commonly encountered significance tests, viz. single sample mean, difference between two independent sample means, single sample proportion, and sample correlation coefficient.

Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

This equally – popular (along with Hays) textbook also contains a clear presentation of the basic notions of statistical power, with special relevance to single - classification ("one-way") analysis of variance.

Wright, B. D. "Sample Free Test Calibration and Person Measurement." *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, New Jersey: Educational Testing Service, 1968.

A most convincing plea for and description of a procedure devised by Rasch for freeing psychological measurement from the particular instrument employed and from previous measures obtained.

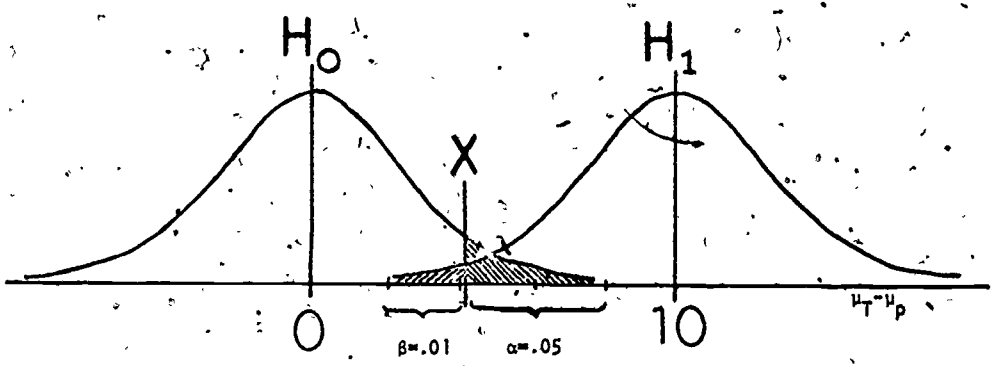
Null (H_0: $\mu_T-\mu_p=0$) and Alternative (H_1: $\mu_T-\mu_p=10$) Sampling Distributions

for Power =.99 (β=.01)

Significance level (α)=.05

Common population variance = 7.66

Sample size = 19



OCCASIONAL PAPERS

1. THE PROBLEM AND PROBLEM DELINEATION TECHNIQUES – William J. Gephart, Phi Delta Kappa. Presented at the Second National Symposium for Professors of Educational Research, sponsored by Phi Delta Kappa, Boulder, Colorado, November 21, 1968. A discussion of the nature of the concept "problem" as related to educational research with a discussion of several techniques useful in problem identification and delineation. $1.00
2. A REVIEW OF INSTRUMENTS DEVELOPED TO BE USED IN THE EVALUATION OF THE ADEQUACY OF REPORTED RESEARCH – Bruce B. Bartos, Phi Delta Kappa & Indiana University. Presented at the Annual Meeting of the American Educational Research Association, February 1969, Los Angeles, California. A brief description and bibliographic annotation of 40 instruments developed to be used in assessing the methodological quality of completed research. $.25
3. PROFILING EDUCATIONAL RESEARCH – William J. Gephart, Phi Delta Kappa, January 1969. The rationale for the development of a methodology profile on completed research to show its strengths and weaknesses. Included are flow charts for profiling the five facets of the research process. $.75
4. APPLICATION OF THE CONVERGENCE TECHNIQUE TO READING – William J. Gephart, Phi Delta Kappa, January 1969. An interim report on a research program planning effort in the field of reading. Free
5. THE CONVERGENCE TECHNIQUE AND READING: A PROGRESS REPORT – William J. Gephart, Phi Delta Kappa. Presented at the Annual Meeting of the International Reading Association, May 2, 1969, Kansas City, Missouri. A second interim report on the planning of a reading research program. Free
6. THE EIGHT GENERAL RESEARCH METHODOLOGIES: A FACET ANALYSIS OF THE RESEARCH PROCESS – William J. Gephart, Phi Delta Kappa, July 14, 1969. The identification and description of general research methods in education through the use of Gutman's facet design and analysis technique. It also details the procedures for the Gutman technique. This paper was printed in the procedings of the Warsaw, Poland Congress of the International Association for the Advancement of Educational Research. Free
7. PROFILING INSTRUCTIONAL PACKAGE – William J. Gephart & Bruce B. Bartos, Phi Delta Kappa, August, 1969. An instruction text to assist individuals with no prior research training in the use of research profiling flow charts to assess the methodological adequacy of completed research. $1.00
8. EDUCATIONAL KNOWLEDGE USE – Gene V Glass, Laboratory of Educational Research, University of Colorado. An analysis of the availability and use of empirically based information in education. $.50
9. MEASUREMENT AND RESEARCH IN THE SERVICE OF EDUCATION – Warren G. Findley, Research and Development Center in Educational Stimulation, University of Georgia. Originally presented as an invited address at the annual meeting of the American Educational Research Association, this paper uses an historical perspective to examine the role of measurement and research in education. $.75

10. THE EDUCATIONAL CATALYST: AN IMPERATIVE FOR TODAY – Joe H. Ward, Jr., Reeve Love, George M. Higginson, Southwest Educational Development Laboratory, Austin, Texas, July, 1971. An analysis of the problems involved in the process of change and improvement of the practice of education. This paper poses a new professional speciality for the facilitation of empirically based educational improvements. $1.00
11. DISSERTATIONS YOU MAY WANT TO SEE – William J. Gephart, Phi Delta Kappa, 1970. A collection of dissertations done in 1969 which focus on research training. $.25
12. THE DOCTORATE IN EDUCATION IN CANADA – Neville L. Robertson, Commission on Higher Education, Phi Delta Kappa, 1971. An analysis of the institutions offering the doctorate in education in Canada. This paper is a companion piece to a larger study of similar institutions in the United States. $.75
13. THE IMPORTANCE OF STATISTICAL POWER IN EDUCATIONAL RESEARCH – John K. Miller, Thomas R. Knapp, University of Rochester. When an educational experiment results in non-significant differences can it be said that no difference exists? This paper discussed the concept that must be attended to IF that question is to be answered. It also details the procedure for determination of sample size needed in an experiment. $1.25